Český webový archiv v roce 2015

Objem dat na internetu se stále zvyšuje a množství těchto dat se mění a zaniká. Český webový archiv zajišťuje výběr, archivaci, dlouhodobou ochranu a zpřístupnění národních webových zdrojů. V současné době význam webových archivů roste a zájem o data z webových archivů mají i vědci. Cílem článku je poskytnout přehled o fungování českého webového archivu a informovat o možnostech jeho využití. Nastíněna je také situace webových archivů v zahraničí.

Úvod

Český webový archiv se nazývá Webarchiv a spravuje jej Národní knihovna České republiky. V současné době je Webarchiv integrální součástí knihovny a na starosti jej má Oddělení archivace webu, které spadá pod Odbor digitálních fondů. Počátky českého webového archivu sahají až do roku 2000, kdy vznikl v rámci společného projektu Národní knihovny ČR, Moravské zemské knihovny v Brně a Ústavu výpočetní techniky Masarykovy univerzity v Brně. Tímto se Webarchiv řadí mezi nejstarší webové archivy v Evropě.

Cílem Webarchivu bylo od počátku archivovat národní (český) web, zajistit jeho dlouhodobou ochranu a zpřístupnit archivované verze uživatelům. Již v roce 2000 se mluvilo o potřebě archivovat web, a to hlavně z důvodu přesunu publikační činnosti na internet. Nicméně v roce 2015 víme, že internet, potažmo web, hraje roli napříč celou současnou společností, ať už se jedná o její sociální, politickou nebo kulturní činnost. Webový obsah se bezesporu stal součástí kulturního dědictví každého národa a otázkou není proč jej archivovat, ale jakým způsobem.

Problémem a zároveň největší výzvou v rámci archivace webového obsahu je jeho dynamika. Webový obsah se mění, mizí, přesouvá se, a to velmi rychle a nekontrolovaně. Všechny webové archivy čelí mnoha výzvám, neboť technický proces sklízení a řízení přístupu k archivu je obtížný. Nové webové technologie a nové typy obsahů vyžadují neustálou změnu nastavení, úpravy nástrojů a pracovních postupů[1].

Webové archivy v zahraničí

Významnou institucí v oblasti archivace webu je mezinárodní konsorcium IIPC (Internatinal Internet Preservation Consortium), které aktuálně sdružuje přes třicet členů. Mezi členy konsorcia IIPC patří nejen webové archivy národních či jiných knihoven (např. webové archivy knihoven Velké Británie, Francie, Dánska nebo Slovinska), ale také webové archivy univerzit (např. na Stanfordově univerzitě a na Kolumbijské univerzitě), případně webové archivy, které jsou neziskovými organizacemi (např. Internet Archive nebo Internet Memory Foundation). Od roku 2005 je také Webarchiv stálým členem konsorcia.

---

[1] GOETHALS, Andrea, Clément OURY, David PEARSON, Barbara SIERMAN a Tobias STEINKE. Facing the Challenge of Web Archives Preservation Collaboratively: The Role and Work of the IIPC Preservation Working Group. *D-Lib Magazine* [online]. 2015, **21**(5/6): - [cit. 2015-09-02]. DOI: 10.1045/may2015-goethals. ISSN 1082-9873. Dostupné z: http://www.dlib.org/dlib/may15/goethals/05goethals.html

Konsorcium sdružuje jednotlivé členy a umožňuje jim navázat spolupráci, pořádá každoročně sjezd členů a konferenci, vytváří standardy v oblasti archivace webu. Jednou z nejpodstatnějších činností konsorcia IIPC je vývoj a správa nástrojů pro archivaci. Konsorcium se zabývá zejména vývojem nástrojů k procházení a sklízení dat (Heritrix) a k jejich zobrazení a zpřístupnění (Wayback machine), ale také například nástroje pro charakterizaci souborových formátů (Jhove2) nebo kurátorských nástrojů (např. NetarchiveSuite). Mnoho z vyvíjených nástrojů je licencováno jako open-source nebo pod jinými veřejnými licencemi[2].

Zřejmě nejznámější organizací zabývající se archivací webu je americká nezisková organizace Internet Archive (*http://www.archive.org*). Organizace Internet Archive sbírá archivní data z webových zdrojů již od roku 1996[3]. Na rozdíl od webových archivů národních knihoven má mezinárodní záběr, archivuje stránky z různých zemí a s různými doménami, ale také další typy elektronických dokumentů, jako například hudbu, video nebo počítačové hry. Tyto archivované verze také zpřístupňuje, jelikož na základě americké legislativy lze tato data nekomerčně zpřístupňovat a v případě požadavku vydavatele může vybraná data znepřístupnit[4].

Některé z národních knihoven zavedly tzv. elektronický povinný výtisk, o jehož naplňování se starají především oddělení archivace webu. Definice elektronického povinného výtisku (electronic legal deposit nebo e-deposit) se různí podle právních předpisů jednotlivých zemí, jedná se však zejména o rozšíření stávajícího povinného výtisku na elektronické publikace. Například Francie přijala tento zákon a Národní knihovna Francie je tak na jeho základě oprávněna archivovat všechny typy elektronických online publikací, nejen webové stránky, ale například i streamované vysílání. Tento zákon se ve Francii vztahuje na sběr dokumentů, ne však na jejich online zpřístupnění, archivovaná data Národní knihovny Francie jsou dostupná pouze v budově knihovny[5]. Britská národní knihovna a její webový archiv také sklízí webové zdroje na základě elektronického povinného výtisku, ovšem stejně jako Webarchiv Národní knihovny ČR získává souhlas vydavatelů pro zpřístupnění výběrové části archivu[6].

Akvizice dat

Archivací webu je označován proces, který zahrnuje výběr webových zdrojů k archivaci, stahování dat z těchto zdrojů (sklízení), jejich dlouhodobé uložení a v neposlední řadě zpřístupnění těchto dat. Z technického hlediska funguje procházení webu a stahování dat na podobném principu jako sběr dat roboty (crawlery) internetových vyhledávačů jako je Google. Webarchiv k tomuto sklízení používá mezinárodně vyvíjený nástroj Heritrix. Proces procházení, stahování a zpřístupnění dat je na rozdíl od výběru zdrojů k archivaci převážně automatizovaný, Heritrix prochází stránky podle zadaných parametrů (např. hloubka sklízení – tj. počet dotazů na doménu) a stahuje data do tzv. balíčků speciálního archivního formátu WARC. Nakonec dochází k indexování těchto dat.

Ve Webarchivu provádíme tři typy archivace – tzv. sklizně, sklizně tematické, které jsou zaměřeny na archivaci zdrojů vztahujících se k určitému tématu (např. prezidentské volby), celoplošné sklizně, kdy je cílem zachytit „celý český web", tzn. webové zdroje s doménou .cz a tato sklizeň se provádí automatizovaně a alespoň jednou ročně. Posledním typem jsou tzv. výběrové sbírky, které obsahují zdroje vybrané kurátory Webarchivu nebo navržené vydavateli a návštěvníky Webarchivu. Pro každou tematickou kategorii jsou vybírány zdroje s různou prioritou, zejména tedy webové stránky zastřešujících a odborných institucí, například pro kategorii politických věd jsou to stránky ministerstev, politologických ústavů vysokých škol, politických stran atd.

Při výběru zdroje k archivaci je přihlíženo k jeho obsahové hodnotě (z historického, kulturního nebo uměleckého hlediska), aktuálnosti a jedinečnosti, souvislosti s dalšími zdroji a v neposlední řadě také je zdroj posuzován z technického hlediska, zda jej bude možné archivovat v odpovídající kvalitě. Každý zdroj je posuzován individuálně. Pro registraci a administraci zdrojů pro výběrové sbírky slouží tzv. kurátorské nástroje (curator tools), Webarchiv používá vlastní nástoj vytvořený přímo pro potřeby českého prostředí. V případě, že je zdroj schválen, je zahájena komunikace s vydavatelem za účelem získání souhlasu s archivací jeho zdroje a jeho zpřístupněním.

Zpřístupnění webového archivu pro uživatele

Přístup do Webarchivu je možný z webové adresy *http://www.webarchiv.cz* nebo z terminálů v budově Národní knihovny. Rozdílem v přístupu je rozsah dostupného obsahu, který je daný zněním autorského zákona. Na terminálech v budově jsou dostupná archivovaná data ze všech sklizní, výběrových, tematických, ale i celoplošných. Online přístup je možný pouze k archivovaným verzím zdrojů, které svůj souhlas s archivací vyjádřily smluvně nebo licenčně, tzn. výběrové sklizně.

Na jaře letošního roku jsme spustili nový design našeho webu, který by měl odpovídat současným požadavkům na jednoduchost a funkčnost. Ústředním prvkem nového vzhledu je aktuálně vyhledávací pole, úvodní strana dále obsahuje vybrané zdroje z našeho katalogu výběrových sbírek na aktuální nebo zajímavé téma (např. 600 let od výročí upálení Jana Husa).

---

[2] Tools and software. *International Internet Preservation Consortium: IIPC* [online]. c2012 [cit. 2015-08-23]. Dostupné z: *http://www.netpreserve.org/web-archiving/tools-and-software*

[3] About the Internet Archive. *Internet Archive* [online]. 1996 [cit. 2015-09-02]. Dostupné z: *https://archive.org/about/*

[4] Removing Documents From the Wayback Machine. *Internet Archive* [online]. 1996 [cit. 2015-09-23]. Dostupné z: *http://archive.org/about/exclude.php*

[5] Digital legal deposit: four questions about Web Archiving at the BnF. *Bibliothèque nationale de France: BnF* [online]. 2014 [cit. 2015-08-24]. Dostupné z: *http://www.bnf.fr/en/professionals/digital_legal_deposit/a.digital_legal_deposit_web_archiving.html*

[6] The British Library Collection Development Policy for websites. *The British Library* [online]. 2014 [cit. 2015-08-24]. Dostupné z: *http://www.bl.uk/aboutus/stratpolprog/digi/webarch/bl_collection_development_policy_v3-0.pdf* [2] Tools and software. *International Internet Preservation Consortium: IIPC* [online]. c2012 [cit. 2015-08-23]. Dostupné z: *http://www.netpreserve.org/web-archiving/tools-and-software*

Vyhledávat je možné již z úvodní stránky Webarchivu prostřednictvím vyhledávacího pole, vyhledávat lze pomocí celé URL adresy nebo její části (např. www.nkp.cz, nkp.cz) nebo pomocí klíčových slov (např. vědecká knihovna). Vyhledávač prohledává kromě URL adres i názvy stránek a také dokáže vyhledávat v anotacích a klíčových slovech přiřazených při katalogizaci. Rádi bychom také v budoucnu umožnili i fulltextové vyhledávání, které je ovšem v současné době stále technicky náročné, jelikož je třeba indexovat kromě stávajícího obsahu na webu (jako to dělají webové vyhledávače jako je Google) také historické archivované verze, kdy každý měsíc přibývá velké množství dat. Webarchiv umožňuje také prostřednictvím rozšířeného vyhledávání, vyhledat archivované verze požadované webové stránky v rozmezí zvolených let.

Druhým způsobem vyhledávání je prohlížení (browsing) záznamů zdrojů, které jsou součástí výběrových sbírek a jejich archivované verze jsou dostupné online.

V předchozí verzi stránek bylo možné prohlížet stránky i na základě jejich vydavatele, to však nebylo uživateli příliš používáno, proto jsme se rozhodli prohlížení zjednodušit pouze na prohlížení tematických kategorií, tzv. katalog stránek. Tento tematický katalog se řídí metodou konspektu, která je používána pro celý fond Národní knihovny ČR. Metoda konspektu obsahuje věcné třídění dokumentů do 24 tematických kategorií (např. biologické vědy, sociologie, zemědělství aj.), které se dále dělí na podkategorie[7]. Katalog stránek tedy přináší přehled všech zdrojů, které pravidelně archivujeme v rámci výběrových sklizní a jejich archivované verze můžeme na základě souhlasu vydavatele zpřístupňovat online. V katalogu je možné přepínat mezi dvěma typy zobrazení, vizuálním (které obsahuje grafické náhledy stránek) a textovým (které obsahuje anotaci a klíčová slova přiřazená ke zdroji).

Posledním způsobem, jak vyhledávat ve zdrojích zařazených do Webarchivu je vyhledávání v katalogu Národní knihovny. Pro každý zdroj zařazený do výběrové kolekce je vytvářen katalogizační záznam podle katalogizačních pravidel a je tak snadno dohledatelný v rámci celého fondu knihovny.

Webarchiv používá pro prezentaci archivovaných dat nástroj s otevřeným zdrojovým kódem (open-source) OpenWayback, který je vyvíjen a spravován komunitou členů IIPC a používá ho většina webových archivů na světě, například i webový archiv Britské knihovny nebo Standfordské univerzity. Tento nástroj umožňuje prohlížet data uložená ve WARC balíčcích v uživatelsky přívětivém prostředí, kdy uživatel prohlíží stránky v takové podobě, v jaké se nacházely v dané době na internetu. Podoba těchto archivovaných stránek má však svá omezení, nemusí být sklizena všechna data nebo je OpenWayback nedokáže z technických důvodů všechna zobrazit. Zobrazená data tak mohou být nekompletní, ale vždy odpovídají tomu, co se na stránkách v minulosti nacházelo, obsah není v žádném případě změněn nebo upraven.

Uživatelské prostředí OpenWayback-u Webarchivu je jednoduché s intuitivním ovládáním. Po výběru z tematického katalogu nebo vyhledání požadovaného zdroje zobrazí časovou osu, na které je zobrazen přehled archivovaných verzí v čase. Lze tak procházet jednotlivými archivovanými verzemi stránek a přepínat mezi nimi.

Uživatelská základna českého webového archivu

Webarchiv jako český webový archiv je určen široké veřejnosti, stejně jako ostatní fondy Národní knihovny ČR. Bohužel, z důvodu omezení autorským zákonem, není možné všechna archivovaná data zpřístupnit online, přístup k celému archivu je tak zatím možný pouze z terminálů v budově knihovny. I přesto je však část archivovaných dat dostupná online prostřednictvím webových stránek Webarchivu. Uživatelé Webarchivu se dají rozdělit zhruba do tří skupin na základě jejich informačních potřeb. Nejpočetnější skupinou jsou individuální uživatelé, kteří přichází do archivu s určitými jednotlivými požadavky, které naplňují zpravidla vlastním prohlížením archivních dat.

Jako další skupinu uživatelů je možné vymezit institucionální uživatele, jako jsou například soudy, úřady průmyslového vlastnictví, policie, výzkumné ústavy a další. Tyto uživatelé využívají data z webového archivu pro svoji profesní činnost a vyžadují potvrzení jejich autenticity. Příkladem může být požadavek patentového úřadu na informace z webového archivu k řešení sporu o práva k technickému řešení, kdy informace z webových archivů mohou prokázat zveřejnění daného řešení v určitém čase. Tyto požadavky jsou zatím poměrně ojedinělé, ale lze předpokládat jejich nárůst.

Další skupinou uživatelů, která bude narůstat do budoucna, jsou výzkumníci a vědci. *„Současným trendem v oblasti archivace webu je rostoucí význam a využití rozsáhlých souborů dat získaných z webových archivu Tyto tzv. big data mohou sloužit pro zkoumání jazyka, technologie, historie nebo dalších oblastí[8].”* Toto potvrzují zkušenosti světových webových archivů, data z webových archivů jsou stále více využívána ve vědeckém zkoumání a výzkumníci a vědci se také stále častěji zapojují k zájmu o analýzy dat z webových archivů.

Ke sledování návštěvnosti webových stránek Webarchivu používáme volně dostupný nástroj Google Analytics. Průměrná návštěvnost webových stránek Webarchivu je mezi jedním až dvěma sty návštěvníků za den. Více než 80 % návštěvníků stránek jsou ovšem noví návštěvníci, kteří navštívili tento web poprvé. Jak lze předpokládat z charakteru webového archivu, většina návštěv webu je z České republiky (přes 80 %), dále pak ze Slovenska (5 %), Německa (2 %) nebo Spojených států (1,5 %)[9].

Závěr

V dnešní době význam a využití webových archivů stále vzrůstá a webové archivy budou stále relevantnější zejména pro získávání historických dat. Nejde jen o individuální informační požadavky, jako prohlížení jednotlivých archivních verzí webo-

[7] Metoda konspektu. *Fondy Národní knihovny ČR* [online]. Praha: Národní knihovna ČR, c2002 [cit. 2015-08-31]. Dostupné z: *http://konspekt.nkp.cz/konspekt.phtml*

[8] KVASNICA, Jaroslav a Barbora BJAČKOVÁ. Profilování kolekce a stanovení určené skupiny WebArchivu Národní knihovny ČR. *ProInflow* [online]. 2014, **6**(No. 1) [cit. 2015-09-02]. ISSN 1804–2406. Dostupné z: http://www.phil.muni.cz/journals/index.php/proinflow/article/view/942

[9] Data ze srpna 2015

vých stránek, ale i práce s velkými soubory dat (tzv. big data). Již dnes se novodobí historici potýkají s nedostatkem původních informací z doby, kdy již existoval internet (90-00 léta), neboť původní webové zdroje zanikly nebo byly změněny. Bez webových archivů by tyto informace přestaly existovat úplně.

Proto je jedním z dlouhodobých cílů Webarchivu vybudování platformy, která by umožnila badatelům přístup k datovým setům, které budou vytvořeny na základě jejich požadavků. Taková platforma si ovšem žádá velké investice do technické infrastruktury, která si musí umět poradit s velkými daty v řádu stovek terabajtů. U zahraničních webových archivů se osvědčilo úložiště na bázi Apache Hadoop, které právě umožňuje zpracování velkého objemu nestruktorovaných dat. Apache Hadoop je využíván i mezi členy konsorcia IIPC a tato komunita pro něj vyvíjí i speciální nástroje, které jsou nutné pro zpracování specifických dat, které mají webové archivy.

Použitá literatura

About the Internet Archive. *Internet Archive* [online]. 1996 [cit. 2015-09-02]. Dostupné z: *https://archive.org/about/*

Digital legal deposit: four questions about Web Archiving at the BnF. *Bibliothèque nationale de France: BnF* [online]. 2014 [cit. 2015-08-24]. Dostupné z:*http://www.bnf.fr/en/professionals/digital_legal_deposit/a.digital_legal_deposit_web_archiving.html*

GOETHALS, Andrea, Clément OURY, David PEARSON, Barbara SIERMAN a Tobias STEINKE. Facing the Challenge of Web Archives Preservation Collaboratively: The Role and Work of the IIPC Preservation Working Group. *D-Lib Magazine* [online]. 2015, **21**(5/6): - [cit. 2015-09-02]. DOI: 10.1045/may2015-goethals. ISSN 1082-9873. Dostupné z: *http://www.dlib.org/dlib/may15/goethals/05goethals.html*

KVASNICA, Jaroslav a Barbora BJAČKOVÁ. Profilování kolekce a stanovení určené skupiny WebArchivu Národní knihovny ČR. *ProInflow* [online]. 2014, **6**(No. 1) [cit. 2015-09-02]. ISSN 1804–2406. Dostupné z: *http://www.phil.muni.cz/journals/index.php/proinflow/article/view/942*

Metoda konspektu. *Fondy Národní knihovny ČR* [online]. Praha: Národní knihovna ČR, c2002 [cit. 2015-08-31]. Dostupné z: *http://konspekt.nkp.cz/konspekt.phtml*

Removing Documents From the Wayback Machine. *Internet Archive* [online]. 1996 [cit. 2015-09-23]. Dostupné z: *http://archive.org/about/exclude.php*

The British Library Collection Development Policy for websites. *The British Library* [online]. 2014 [cit. 2015-08-24]. Dostupné *z: http://www.bl.uk/aboutus/stratpolprog/digi/webarch/bl_collection_development_policy_v3-0.pdf*

Tools and software. *International Internet Preservation Consortium: IIPC* [online]. c2012 [cit. 2015-08-23]. Dostupné z: *http://www.netpreserve.org/web-archiving/tools-and-software*

Mgr. Barbora Rudišinová
barbora.rudisonova@nkp.cz

Bc. Jaroslav Kvasnica
jaroslav.kvasnica@nkp.cz

(Národní knihovna ČR)